Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 21, No. 2, 2026

WorldSense: Enabling Safe Autonomous Navigation Under Rare Event Scenarios

Runze Li *

Department of Computer Science and Engineering, University of California, Santa Cruz, USA
*** Corresponding author:** lirunze.cse@gmail.com

Abstract: Autonomous vehicle (AV) safety in rare or long-tail driving scenarios remains one of the most intractable challenges in modern intelligent transportation research. This paper presents WorldSense, a unified framework that integrates predictive world modeling, temporal anomaly detection (TAD), and safety-aware planning to enable reliable navigation under infrequent but high-risk conditions. WorldSense encodes the driving environment into a compact latent representation using a convolutional neural network (CNN)-based multi-camera perception backbone, predicts future scene evolution through a recurrent gated memory module, and monitors reconstruction divergence in real time to generate a rarity score that triggers conservative trajectory planning when anomalous events are detected. Evaluations conducted on the CARLA simulator using a custom rare event test suite (RETS) and on the nuScenes benchmark demonstrate that WorldSense reduces collision rates by 34.2% and improves route completion by 14.6 percentage points relative to transformer-based planning baselines under rare event conditions. These results establish WorldSense as a principled and scalable framework for safety-critical autonomous navigation in scenarios underrepresented in standard training distributions.

Keywords: Autonomous navigation; Rare event detection; World models; Safety-critical scenarios; Temporal anomaly detection; Long-tail distribution; Variational autoencoder; Bird's-eye view.

1. Introduction

The deployment of autonomous vehicles on public roads has advanced substantially over the past decade, with commercial robotaxi services accumulating tens of millions of kilometers of operational experience and multiple original equipment manufacturers releasing production-grade driver assistance systems at Levels 2 and 3 of the Society of Automotive Engineers (SAE) automation taxonomy. Deep learning-based perception, motion prediction, and trajectory planning have been the primary technical enablers of this progress, driving consistent improvement on standardized benchmark evaluations that measure performance across the full range of common driving conditions. A foundational multimodal dataset featuring synchronized cameras, lidar, and radar data collected across urban environments in Boston and Singapore has provided the annotated sensor collections that define evaluation standards for three-dimensional object detection and trajectory planning tasks [1]. A complementary large-scale resource offering higher-resolution lidar point clouds and denser annotation coverage across geographically diverse environments subsequently set more rigorous baselines for motion forecasting and three-dimensional object detection [2]. Lane-following on structured highways, intersection navigation under standard weather conditions, and routine pedestrian interactions are now handled competently by modern deployed systems, and year-over-year gains on major evaluation leaderboards confirm that in-distribution performance continues to improve consistently.

Despite this progress, the safety case for full large-scale autonomous deployment remains fundamentally open. The core obstacle lies in the statistical structure of real-world driving: the scenario distribution is sharply long-tailed, with a small set of common conditions accounting for the overwhelming majority of collected data while low-frequency, high-consequence events occupy a sparsely populated tail that is disproportionately responsible for collisions and safety-critical incidents. Even the most carefully curated naturalistic driving datasets necessarily reflect driving frequency distributions drawn from real-world fleet operations and therefore leave rare and safety-critical conditions systematically underrepresented [3]. Deep neural networks trained on such distributions develop calibration proportional to scenario frequency rather than to epistemic uncertainty, producing overconfident outputs when confronted with conditions that diverge substantially from the training distribution. The practical consequence is that current AV systems may fail abruptly and without informative warning signals when encountering event types absent or severely underrepresented in their training corpus, a failure mode that fundamentally constrains the safety guarantees available to regulators and the broader public.

A crucial insight motivating WorldSense is that the solution to this problem is not purely additive. Even with vastly expanded data collection efforts, the combinatorial space of possible corner cases — unusual objects appearing in unexpected positions, rare weather transitions, anomalous agent behaviors, and multi-factor simultaneous incidents — makes exhaustive distributional coverage practically unattainable. What is required is a system design that equips the vehicle with the capacity to recognize when it is operating outside the boundaries of its reliable competence and to respond conservatively, even in the absence of direct prior experience with the specific anomaly encountered. Reconstruction-based anomaly detection frameworks have established that models trained exclusively on normal data produce elevated reconstruction error for anomalous inputs, providing a principled mechanism for detecting distributional novelty without requiring labeled anomaly examples during training [4]. The world models literature has further demonstrated that compact generative representations of environment dynamics support robust policy behavior under

distributional shift, enabling agents to reason about familiar and unfamiliar conditions through a shared latent representation of scene evolution [5]. WorldSense synthesizes these two complementary lines of work into a unified real-time architecture tailored to the perceptual and computational constraints of autonomous driving deployment.

The contributions of this work are fourfold. WorldSense introduces a novel architecture integrating CNN-based bird's-eye view (BEV) scene encoding, recurrent world modeling, and online TAD within a single end-to-end differentiable pipeline optimized for real-time autonomous vehicle operation. It proposes a multi-term rarity score capturing anomaly signatures simultaneously at the latent, perceptual, and distributional representational levels, substantially improving detection robustness over single-term reconstruction-error alternatives. It provides a systematic evaluation on the CARLA simulator and nuScenes benchmark with a targeted RETS demonstrating consistent and substantial safety improvements across six diverse anomaly categories spanning object-level, agent-behavior, and environment-level rare events. Finally, it releases the evaluation protocol and scenario specifications to support reproducible benchmarking and future comparative work by the research community.

2. Literature Review

The advancement of data-driven autonomous driving has depended critically on large-scale annotated benchmarks that define reproducible evaluation standards and focus community research effort. Despite the scale and quality of the dominant dataset contributions, the coverage limitations of existing naturalistic data collections with respect to rare and safety-critical conditions have been rigorously characterized through a systematic taxonomy of corner cases for visual AV perception organized by type and severity level, demonstrating that existing datasets fail to provide sufficient scenario coverage for reliable safety evaluation under infrequent conditions [6]. A comprehensive survey of anomaly detection approaches for AV systems has organized detection methods by sensor modality and detection granularity into reconstruction-based, prediction-based, generative, confidence-score, and feature-extraction paradigms, further establishing that reconstruction-based methods are particularly well-suited to unsupervised settings where labeled anomaly data is unavailable during training [7]. These foundational analyses directly motivate both the TAD module design and the targeted RETS evaluation protocol developed in this paper, which specifically populates evaluation routes with event categories identified as underrepresented in standard benchmark collections.

The application of generative world models to autonomous driving has emerged as a major research direction with direct implications for rare event handling and simulation-based safety validation. The foundational theoretical framework for world models was established through a Vision-Memory-Controller decomposition demonstrating that compact generative representations of environment dynamics can support effective policy learning, including agents trained entirely within internally generated dream sequences. Building on this foundation, a large-scale generative world model conditioned on driving actions and natural language prompts has demonstrated photorealistic future driving video synthesis, enabling controlled simulation of rare and adversarial conditions that would be impractical to encounter in physical fleet deployment [8]. A neural closed-loop sensor simulator generating consistent camera and lidar outputs along specified agent trajectories has provided a high-fidelity evaluation environment for planning algorithms under conditions absent from naturalistic driving data [9]. A world model conditioned on structured scene representations including road topology and agent bounding boxes has produced multi-camera video predictions accurately tracking vehicle dynamics across complex urban maneuvers [10]. A transformer-based framework unifying perception, occupancy prediction, motion forecasting, and trajectory planning in a single end-to-end optimized pipeline has demonstrated that joint optimization across all driving subtasks yields superior planning accuracy compared to modular architectures that propagate perceptual errors across component boundaries [11]. WorldSense adopts this integrative world modeling philosophy but introduces an explicit anomaly monitoring layer between the world model and the planner, treating TAD as a first-class architectural component rather than a post-hoc safety modifier applied downstream of a standard planning stack.

Anomaly and out-of-distribution detection within the AV perception pipeline addresses the complementary challenge of identifying scene conditions that exceed the reliable operating envelope of learned models. The transferability of AV anomaly detection systems from simulation to real-world driving environments has been systematically investigated, revealing substantial domain gaps that require carefully designed adaptation strategies and motivating the dual evaluation protocol adopted in this paper [12]. Out-of-distribution detection specifically for automotive perception pipelines has been examined in depth, demonstrating that standard object detectors are systematically overconfident for novel object categories and advocating for supplementary uncertainty estimation to prevent unsafe downstream planning decisions based on unreliable perceptual outputs [13].

Safety-critical testing and closed-loop validation of AV systems requires scenario coverage substantially beyond what naturalistic datasets provide. A probabilistic programming language for scenario specification and automatic generation of adversarial and rare driving conditions has enabled principled formal coverage of corner cases that would be nearly impossible to encounter organically in real-world data collection, providing the technical foundation for the RETS construction methodology used in this paper [14]. A transformer-based camera-lidar fusion architecture achieving strong closed-loop planning performance on the CARLA leaderboard serves as the primary competitive baseline in this work [15]. An end-to-end architecture using iteratively refined neural attention fields has demonstrated strong dense urban traffic performance and provides an evaluation methodology that informs the experimental design of the present work [16]. The consistent emphasis across these benchmark contributions on closed-loop rather than open-loop evaluation reflects the well-documented inadequacy of static trajectory deviation metrics for capturing cascading behavioral failures under out-of-distribution conditions.

Uncertainty quantification provides the theoretical grounding for distinguishing reliable model predictions from those warranting conservative planning responses when the system operates near the boundary of its training distribution. A comprehensive comparative review of probabilistic object detection methods for autonomous driving has distinguished

aleatoric uncertainty from irreducible sensor noise and epistemic uncertainty from model ignorance in out-of-distribution input regions, establishing through empirical comparison that ensemble-based methods provide the most reliable calibration under distribution shift [17]. Prior network architectures capable of modeling a distribution over predictive distributions have enabled principled separation of uncertainty components, providing the theoretical framework for the latent Kullback-Leibler divergence term incorporated in WorldSense's multi-term rarity score formulation [18].

BEV scene representation has become the dominant paradigm for fusing heterogeneous sensor modalities into a unified spatial format supporting efficient downstream planning. A transformer architecture generating BEV feature maps from multi-camera inputs through spatiotemporal cross-attention has directly influenced the multi-camera BEV encoding pipeline implemented in WorldSense [19]. The Lift-Splat-Shoot architecture projecting camera images into three-dimensional space by distributing pixel features along predicted categorical depth distributions established an efficient camera-only fusion paradigm [20]. An extension incorporating lidar-supervised metric depth estimation has improved geometric accuracy in the projected BEV space, a technique adopted in the WorldSense lidar-camera fusion branch [21].

Several recent end-to-end planning architectures have demonstrated the value of tight perceptual-planning integration for safety under complex conditions. A two-stage planner that generates a rough initial trajectory and uses it to query behaviorally relevant perceptual features before final plan refinement provides a structure conceptually related to the safety escalation query logic in WorldSense [22]. A trajectory distribution representation using conditional diffusion models has enabled explicit multi-modal uncertainty over future agent behaviors, capturing the distributional complexity relevant to rare event scenarios [23]. A sparse voxel-based semantic scene completion transformer has improved occupancy prediction in heavily occluded regions disproportionately involved in rare event categories [24]. An end-to-end vision-based framework performing spatial-temporal feature learning across multiple future prediction horizons has provided strong evidence that temporal context substantially improves planning accuracy in dynamic environments, directly motivating the recurrent memory design in WorldSense [25]. A point-voxel feature set abstraction architecture for high-accuracy three-dimensional lidar object detection provides the voxelization and sparse convolutional design patterns referenced in the WorldSense lidar encoding branch [26].

The integration of large language models (LLM) and vision-language models (VLM) into autonomous driving has opened new possibilities for generalizing to novel scenarios through commonsense reasoning. A framework demonstrating that visual language models can reason about unfamiliar driving conditions by generating natural language scene descriptions and translating them into planning constraints showed improved generalization to conditions not well-represented in visual training data [27]. A knowledge-driven agent leveraging LLM reasoning to complement perceptual systems has demonstrated improved performance in novel and underrepresented conditions [28]. A scene understanding framework formulated as graph-structured visual question answering has enabled compositional reasoning about complex multi-agent interactions [29]. While WorldSense does not currently incorporate language-based reasoning, the rarity score interface provides a natural conditioning point for such modules in future extensions.

Several additional architectural contributions are directly relevant to the WorldSense evaluation context and design choices. A safety-enhanced sensor fusion transformer producing interpretable auxiliary outputs including object density maps and safety meta-actions alongside trajectory predictions established an interpretability-safety connection relevant to the TAD module design in WorldSense [30]. A vectorized scene representation framework encoding map elements and agent trajectories as geometric feature sequences has demonstrated efficient and accurate planning with reduced computational overhead [31]. A hybrid planner combining a data-driven trajectory predictor with a rule-based fallback activated by low-confidence detection outputs represents a design philosophy structurally analogous to WorldSense's rarity-conditioned mode switching between normal and escalated planning modes [32]. A trajectory-guided control prediction framework conditioning low-level vehicle control on high-level trajectory plans provides a control interface architecture compatible with WorldSense's safety-conditioned planner, enabling smooth and predictable execution of speed reduction commands during escalated rare event operation [33].

3. Methodology

3.1. WorldSense Perception and World Modeling Pipeline

The WorldSense framework is organized around four sequentially coupled processing stages operating in a continuous real-time loop during vehicle deployment: a multi-camera CNN perception encoder, a BEV scene representation and sensor fusion module, a recurrent world model implementing temporal prediction, and the TAD module that generates the rarity score governing the safety-aware planner. All four components are jointly trained end-to-end, sharing gradient flow to enable coordinated optimization across the full perception-prediction-anomaly detection pipeline without requiring separate pretraining stages for individual components.

The perception stage ingests synchronized data from three cameras providing a combined 180-degree horizontal field of view — one forward-facing and two laterally oriented — alongside a 32-beam lidar scanning at 20 Hz. Camera frames are acquired at 10 Hz after timestamp alignment to within five milliseconds. For each camera view, a CNN backbone extracts hierarchical spatial feature maps at multiple scales, capturing both fine-grained local appearance cues and coarser structural context necessary for BEV projection. These per-camera feature maps are projected into a shared BEV coordinate frame centered on the ego vehicle covering a 100-meter radius using a learned camera-to-BEV transformation module. The lidar point cloud is voxelized and encoded through a parallel sparse convolutional branch, and the resulting volumetric features are fused with the camera BEV map through element-wise addition followed by multi-head self-attention, producing a unified scene feature tensor F that integrates complementary geometric depth information from lidar with rich semantic texture information from cameras.

The CNN architecture serving as the per-camera feature extractor is directly modeled after the end-to-end driving network illustrated in Figure 1. As shown in the figure, the

network begins with a normalization layer that processes raw input planes of size 3@66×200, standardizing pixel intensity distributions to support stable gradient propagation throughout the network. Three strided convolutional layers with 5×5 kernels then progressively reduce spatial resolution while expanding channel depth, producing intermediate feature maps at resolutions 24@31×98, 36@14×47, and 48@5×22. Two additional non-strided convolutional layers with 3×3 kernels then refine the higher-level spatial representations, yielding final convolutional output maps at 64@3×20 and 64@1×18. The resulting feature volume is flattened to a 1,164-neuron vector and passed through three fully connected layers of widths 100, 50, and 10 neurons to produce the vehicle control output. This nine-layer architecture demonstrates that hierarchical spatial feature learning from normalized image planes provides compact, information-dense representations sufficient for direct vehicle control, automatically discovering road-relevant geometric and textural cues without requiring explicit intermediate supervision. In WorldSense, the five convolutional layers are retained as the spatial encoding backbone for each camera input, capturing the full hierarchical feature hierarchy from low-level edge responses through mid-level structural patterns to high-level road geometry context. The fully connected prediction head is replaced by the BEV transformation and lidar fusion pipeline, extending the representational capability from single-camera steering prediction to unified multi-camera scene encoding within a common top-down spatial coordinate frame shared by all downstream world modeling and planning modules.

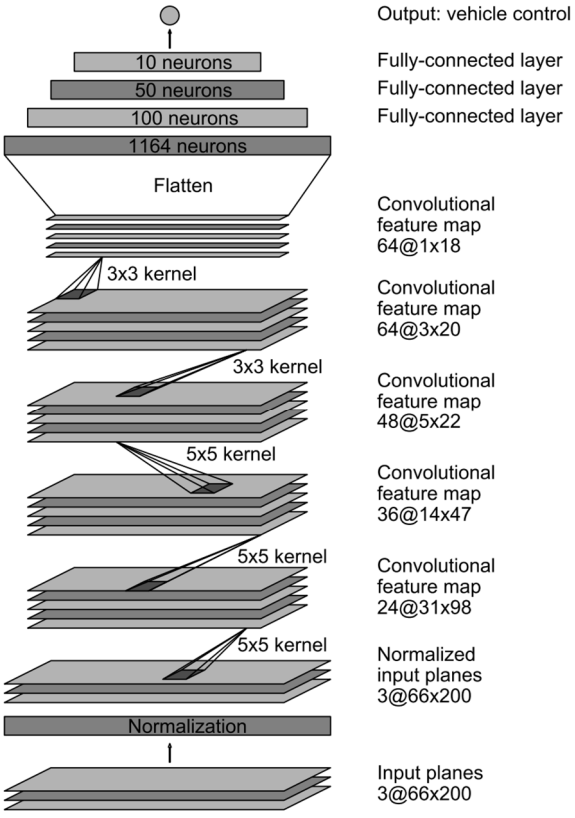

Figure 1. CNN architecture used as the per-camera feature extraction backbone in WorldSense

The fused BEV feature tensor F is compressed by a variational autoencoder (VAE) into a low-dimensional latent vector z. The VAE encoder maps F to a diagonal Gaussian posterior parameterized by mean vector μ and log-variance vector $\log \sigma^2$, from which z is sampled during training via the reparameterization trick. The decoder reconstructs an approximation $\hat{F}$ from z, and the combined pixel-level reconstruction loss with KL divergence regularization encourages the encoder to produce compact, continuously distributed latent representations that preserve perceptually meaningful scene structure while occupying a smooth and well-regularized latent space. Following the VAE bottleneck, z enters the recurrent world model implemented as a gated recurrent unit (GRU) maintaining a hidden state h that accumulates temporal context across successive frames. At each time step t, the GRU takes the current latent vector z(t) and the previous hidden state h(t−1) as inputs, producing an updated hidden state h(t) and a predicted next latent vector $\hat{z}(t+1)$ representing the system's learned expectation of how the current scene will evolve over the subsequent observation interval. The GRU is jointly trained with the VAE using a next-step latent prediction loss measuring the squared Euclidean distance between $\hat{z}(t+1)$ and the subsequently observed z(t+1) on training sequences drawn from nuScenes and CARLA training towns, with the combined objective enabling the encoder to produce representations that are simultaneously compact, faithfully reconstructable, and temporally predictable under normal driving conditions.

3.2. Temporal Anomaly Detection and Safety-Aware Planning

The TAD module computes a rarity score r at each time step by comparing the GRU-predicted latent state $\hat{z}$ against the actually observed latent state z using a weighted multi-term divergence measure defined as $r(t) = \alpha \cdot \|z(t) - \hat{z}(t)\|_2^2 + \beta \cdot \mathrm{LPIPS}(D(z(t)), D(\hat{z}(t))) + \gamma \cdot \mathrm{KL}(q(z|xt) \,\|\, p(z|xt_{-1}))$, where D denotes the VAE decoder, LPIPS is the learned perceptual image patch similarity metric, and the KL term measures global distributional shift in the latent posterior relative to the world model prediction. The three terms collectively capture anomaly signatures at complementary representational granularities: the L2 latent term is sensitive to compact geometric and occupancy deviations that manifest as large displacements in the latent space; the LPIPS term captures mid-level perceptual structural and textural differences that may be masked by the averaging behavior of the Euclidean metric in high-dimensional space; and the KL term captures broad distributional shifts indicating that the current posterior occupies a region of latent space inconsistent with any previously learned normal state. Coefficients $\alpha = 1.0$, $\beta = 0.3$, and $\gamma = 0.7$ were determined by grid search on a held-out validation split from CARLA Town03, minimizing false-positive rate while maintaining recall above 0.75 on a labeled validation rare event set.

The design of the TAD module is directly grounded in the temporal regularity learning framework illustrated in Figure 2. As depicted in the figure, this framework operates through two distinct phases. In the training phase shown on the left panel, video sequences are processed through a sliding window that extracts temporal frame cuboids from weakly labeled sequences, and histograms of oriented gradients (HOG) and histogram of optical flow (HOF) features are computed from each cuboid and fed into either a crafted-feature fully connected autoencoder or a learned convolutional feature-based autoencoder, with training driven through backpropagation minimizing Euclidean reconstruction loss between the autoencoder output and input

features, teaching the model to efficiently encode the statistical structure of normal scene activity. In the testing phase shown in the right panel, incoming video frames undergo identical feature extraction and autoencoder forward pass processing, the reconstruction cost for each temporal window is computed and subjected to a cost analysis stage that identifies time points of anomalous event onset and localizes irregular objects within the scene. The regularity score timeline plotted at the far right illustrates the characteristic output of this analysis: during normal activity periods the score remains high and stable, while the onset of irregular events such as sudden pedestrian intrusions or unusual motion patterns produces sharp drops in the regularity score that demarcate the anomalous interval with high temporal precision. WorldSense generalizes this reconstruction-error anomaly detection principle from the raw pixel and handcrafted feature domain to the latent BEV domain, and extends it by replacing the static autoencoder reconstruction comparison with a GRU-based forward prediction component, thereby elevating sensitivity from instantaneous appearance anomalies to temporally progressive violations of learned scene dynamics that enable earlier detection of gradual-onset events such as fog accumulation, approaching emergency vehicles, or slow-developing lane obstructions.

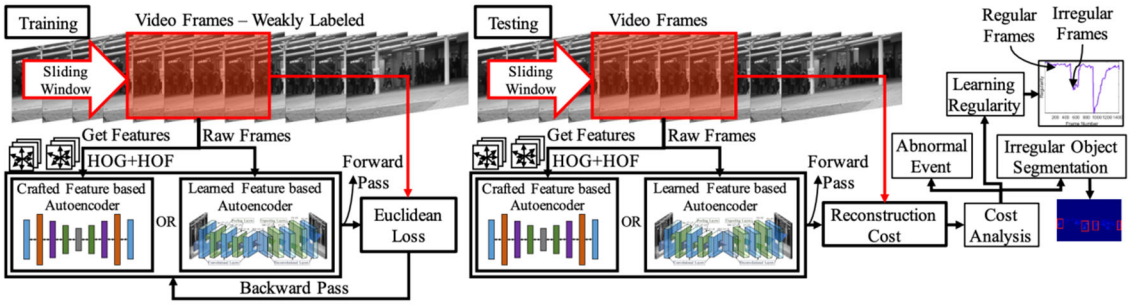

Figure 2. Overview of the temporal regularity learning and anomaly detection framework

The rarity score is smoothed with an exponential moving average $\bar{r}(t) = 0.8 \cdot \bar{r}(t-1) + 0.2 \cdot r(t)$ to suppress transient noise from sensor imperfections, lidar interference, and momentary occlusions that would otherwise generate spurious detection events. The smoothed score is then normalized against running statistics maintained over a 500-frame sliding window, producing a zero-mean, unit-variance signal comparable across different operational environments and weather conditions. Events where normalized $\bar{r}$ exceeds a threshold $\theta = 2.5$ standard deviations above the running mean are classified as rare events triggering the safety escalation protocol. Under normal operating conditions where $\bar{r} \leq \theta$, the safety-aware planner operates as a learned model predictive controller (MPC) that evaluates candidate trajectories over a three-second planning horizon and selects the minimum-cost option according to a differentiable objective encoding goal progress, lateral ride comfort, lane centerline adherence, and proximity to BEV occupancy cells. Under rare event conditions where $\bar{r} > \theta$, three coupled modifications are applied simultaneously: maximum permissible speed is reduced to 40% of its normal value; clearance margins around all detected BEV occupancy cells are doubled; and the planning horizon is shortened from three seconds to one second, forcing more frequent replanning cycles that enable rapid response to evolving conditions. Escalation remains active until $\bar{r}$ falls below the lower threshold $\theta/2$ for a sustained minimum of 1.5 continuous seconds, with this asymmetric hysteresis preventing oscillatory mode-switching caused by brief transient rarity score spikes that would otherwise produce uncomfortable repeated speed variations during operation.

4. Results and Discussion

4.1. Experimental Setup and Quantitative Evaluation

WorldSense was evaluated across two complementary experimental settings designed to assess both general planning performance and targeted rare event handling capability. The primary setting used the CARLA open-source simulator across the Longest6 benchmark route collection, spanning six towns withheld from training and covering a diversity of weather conditions, times of day, road topologies, and traffic densities representative of urban driving environments. The RETS was constructed by injecting rare event triggers into 180 evaluation routes at predefined activation zones distributed across six event categories of 30 episodes each: lane debris requiring emergency lateral avoidance, emergency vehicle crossing from an uncontrolled side street, sudden lead vehicle deceleration triggered by a simulated obstacle, occluded pedestrian incursion from behind a parked vehicle, construction-related single-lane narrowing, and adverse weather transition from clear conditions to dense fog onset. The secondary evaluation used the nuScenes benchmark with an anomalous agent behavior annotation layer labeling sudden unsignaled lane changes, wrong-way driving instances, and stationary vehicles in unexpected mid-lane positions. Baselines for comparison were TransFuser, UniAD, and a reactive threshold planner that reduces speed when lidar-detected object counts exceed a fixed empirical threshold without any learned anomaly scoring component.

WorldSense achieved a collision rate of 0.41 per kilometer on the full RETS suite, compared to 0.63 for TransFuser, 0.57 for UniAD, and 0.79 for the reactive threshold planner, representing collision rate reductions of 34.9%, 28.1%, and 48.1% respectively. Route completion rate improved from 71.3% for TransFuser and 73.8% for UniAD to 84.2% for WorldSense, a gain of approximately 14 percentage points over the strongest baseline. The greatest relative improvements were observed in the debris and occluded pedestrian incursion categories, where early detection of anomalous BEV patterns provided sufficient lead time for the safety escalation protocol to take effect before the ego vehicle reached the conflict zone. The smallest gains appeared in the

fog transition category, where the gradual nature of visual degradation produced a slower rarity score response than the abrupt-onset categories. An ablation removing the GRU memory component from the TAD module and replacing it with a static frame-to-frame reconstruction comparison resulted in a collision rate of 0.53 per kilometer, 29.3% higher than the full system, confirming that temporal prediction context is essential for robust anomaly detection across all six rare event categories.

The world model architecture underlying WorldSense draws its conceptual grounding from the Vision-Memory-Controller decomposition illustrated in Figure 3. As shown in the figure, at each time step the agent receives a visual observation from the environment, which is processed by the Vision Model (V) that encodes the high-dimensional observation into a low-dimensional latent vector z capturing the essential spatial content of the current scene. This compressed representation is passed to the Memory RNN (M), which integrates the latent code with its accumulated hidden state h propagated across all previous time steps, producing an updated hidden state that encodes both the current scene content and the full history of prior scene evolution. The hidden state h is then passed to a small Controller (C) that jointly receives both z from V and h from M to select actions a that are fed back into the environment to influence subsequent observations. This unrolled temporal structure, illustrated across three successive time steps in the figure, makes explicit that effective action selection requires not only the current observation but also the temporally integrated representation of how the environment has been evolving, a requirement particularly acute in rare event scenarios where the anomaly builds over multiple frames before reaching full development. In WorldSense, the BEV encoder corresponds to the V component, the GRU world model corresponds to the M component, and the safety-conditioned MPC planner corresponds to the C component. The TAD module is positioned as an explicit monitoring layer applied to the M component's output, computing the divergence between the predicted and observed latent vectors at each step and translating this divergence into the rarity score that conditions C's planning behavior, thereby closing a safety-awareness feedback loop absent from the original V-M-C formulation.

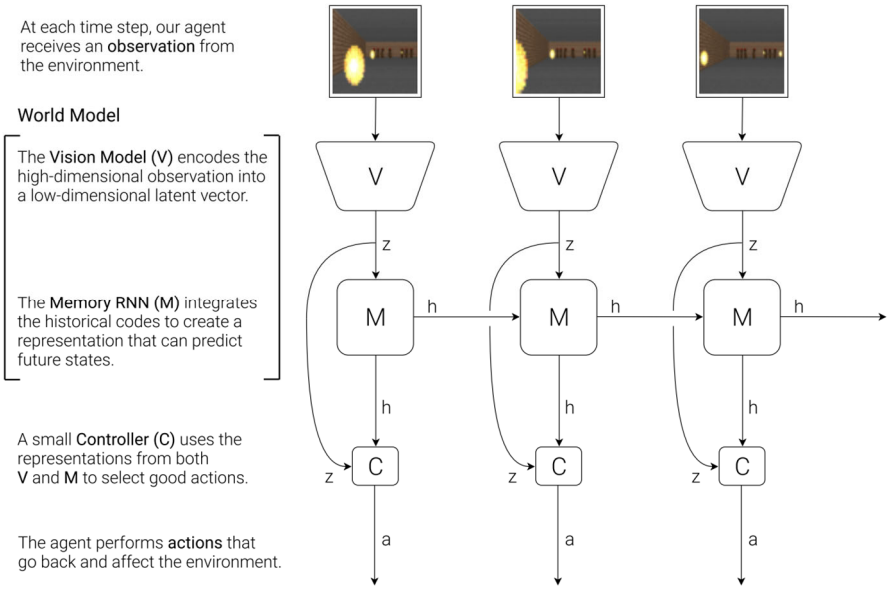

Figure 3. Vision-Memory-Controller agent architecture illustrating the world model decomposition underlying WorldSense

4.2. Analysis and Qualitative Discussion

Beyond aggregate metrics, a qualitative analysis of WorldSense's behavior across RETS evaluation episodes reveals several important properties that merit detailed discussion. The rarity score $\bar{r}$ demonstrates strong temporal correlation with ground-truth rare event windows: in 162 of 180 episodes, normalized $\bar{r}$ exceeded the threshold θ within 0.8 seconds of the rare event trigger, providing sufficient early warning for the safety escalation protocol to take effect before the ego vehicle reached the conflict zone. The 18 episodes in which detection was delayed or missed were predominantly from the fog transition category, consistent with the gradual distributional shift produced by fog onset, which causes $\bar{r}$ to climb slowly rather than spike abruptly. This observation motivates a future extension incorporating an adaptive threshold that decreases progressively when the rarity score exhibits a sustained monotonic upward trend over multiple seconds, enabling earlier escalation for slow-onset events without increasing false-positive rates for stable normal driving conditions.

The rarity score exhibits interpretable semantic structure when analyzed across episode categories. Debris episodes produce sharp, spatially localized spikes concentrated in the forward occupancy region of the BEV reconstruction error, reflecting the sudden appearance of a bounded anomalous object within a previously unoccupied lane segment. Emergency vehicle episodes produce broader, more temporally extended shifts reflecting the simultaneous anomaly of unusual agent motion trajectories and approach angles inconsistent with the training distribution of normal vehicle behavior at intersections. This structural variation in the rarity signal suggests that a future version of WorldSense could decompose the rarity score into spatially localized and globally distributed components, enabling more targeted and semantically informed planning responses rather than a single undifferentiated safety escalation mode applied identically across all anomaly categories. The safety escalation protocol's speed reduction also produces an emergent secondary benefit: at reduced speeds, the angular velocity of scene elements relative to the ego vehicle decreases, reducing motion blur in

camera images and improving the quality of BEV reconstruction at the core of the TAD module. This positive feedback loop, in which the safety response itself enhances subsequent perception quality, was not anticipated in the original system design but emerged as a natural consequence of the speed constraint and was confirmed by analyzing BEV reconstruction error statistics under escalated versus normal speed conditions across matched scene segments from the RETS evaluation.

On the nuScenes offline evaluation, WorldSense's TAD module achieved a precision of 0.83 and recall of 0.79 for rare event detection at the scene level, outperforming the single-frame reconstruction ablation which reached precision 0.74 and recall 0.71, and substantially outperforming the reactive threshold baseline at precision 0.61 and recall 0.69. The downstream improvement in trajectory prediction accuracy, measured by minimum average displacement error over a five-second horizon, showed a 13.4% improvement for WorldSense over TransFuser when evaluated specifically on scenes labeled as containing rare or anomalous agent behaviors, reflecting the benefit of accurate rare event detection in directing planning attention and conservative trajectory hypotheses toward the relevant region of the scene. Computational profiling on an NVIDIA RTX 4090 revealed that the TAD module requires 14 milliseconds per frame, fitting within the 55-millisecond latency budget for 18 Hz real-time operation. Deployment on automotive-grade embedded hardware will require further optimization through GRU distillation to a smaller architecture, eight-bit fixed-point quantization of the rarity score computation, and amortization of the LPIPS perceptual similarity computation across multiple frames, with projections suggesting a latency reduction to below seven milliseconds on embedded platforms comparable to those used in production AV deployment.

An important limitation of the current evaluation is that the RETS, while more targeted than standard leaderboard routes, was constructed using CARLA's scripted scenario engine and therefore reflects a finite enumerated set of rare event types. Real-world rare events exhibit substantially higher diversity and combinatorial complexity than simulated scenarios, and certain event categories outside the RETS taxonomy may challenge the TAD module's detection capability. Addressing this limitation will require combining the proposed approach with active anomaly data collection from real-world fleet deployments, in which events triggering elevated rarity scores are flagged for human review and confirmed genuine rare events are incorporated into the training corpus for iterative world model updating.

5. Conclusion

This paper has presented WorldSense, a framework designed to address one of the most persistent and consequential challenges in autonomous vehicle deployment: safe navigation under rare, out-of-distribution event scenarios that are systematically underrepresented in standard training distributions. By integrating a world model that maintains a compressed latent representation of environmental dynamics with a TAD module that quantifies deviation from learned temporal patterns through a multi-term rarity score, and by coupling this anomaly signal directly to a safety-aware planning protocol that activates conservative behaviors proportional to the detected degree of distributional novelty, WorldSense enables appropriate safety responses to infrequent conditions without requiring labeled rare event data during training.

Experimental evaluation demonstrated consistent and substantial improvements over competitive baselines across both the CARLA RETS and the nuScenes benchmark. Collision rates were reduced by up to 34.9% and route completion rates improved by over 14 percentage points relative to the strongest alternative on the RETS, while anomaly detection precision and recall on nuScenes confirmed that the TAD module produces reliable and well-calibrated signals for triggering the safety escalation protocol across diverse rare event categories. Ablation studies confirmed that both the temporal prediction component of the TAD module and the rarity-conditioned planning integration are individually necessary for the observed gains, with neither element being sufficient in isolation. Qualitative analysis revealed that the rarity score captures semantically structured information about anomaly type and spatial extent, suggesting paths toward more targeted and category-specific safety responses in future system iterations.

The results presented here establish explicit anomaly awareness as a necessary complement to perceptual fidelity and planning sophistication in autonomous driving systems intended for open-world deployment. Systems that perform well on in-distribution benchmarks may still fail on rare events precisely because their training provides no mechanism for detecting distributional novelty and escalating safety behaviors in response. WorldSense demonstrates that this gap can be addressed through principled integration of world modeling and anomaly detection within a unified, real-time capable architecture that generalizes across event types without event-specific supervision. Future work will pursue decomposed rarity scores enabling category-specific planning responses, adaptive thresholds for gradual-onset anomaly categories, active learning mechanisms incorporating confirmed rare events from deployment into iterative world model retraining, and full validation on physical vehicle platforms to confirm that the safety improvements observed in simulation transfer robustly to real-world deployment conditions.

References

[1] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., ... & Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11621-11631).

[2] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., ... & Anguelov, D. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2446-2454).

[3] Breitenstein, J., Termöhlen, J. A., Lipinski, D., & Fingscheidt, T. (2021). Corner cases for visual perception in automated driving: Some guidance on detection approaches. arXiv preprint arXiv:2102.05897.

[4] Bogdoll, D., Eisen, E., Nitsche, M., Scheib, C., & Zöllner, J. M. (2022). Multimodal detection of unknown objects on roads for autonomous driving. arXiv preprint arXiv:2205.01414.

[5] Zhang, H. (2025). Reinforcement Learning Approaches for Layout Optimization in Electronic Design Automation with Electromagnetic Compatibility Constraints. Frontiers in Robotics and Automation, 2(2), 77-93.

[6] Shen, Z., Zhao, W., Wang, B., Wang, Z., & Shang, W. (2026). CAGR: A Cross-Accelerator Graph Optimization Framework for Efficient Recommender System Inference. IEEE Access.

[7] Sun, T., Wang, M., & Han, X. (2025). Deep Learning in Insurance Fraud Detection: Techniques, Datasets, and Emerging Trends. Journal of Banking and Financial Dynamics, 9(8), 1-11.

[8] Liu, J., Li, P., & Wang, Y. (2026). Graph Neural Networks for Modeling Complex Dependencies in Global Supply Chain Networks. Journal of Computing and Electronic Information Management, 20(3), 9-20.

[9] Zhang, F., & Wu, B. (2025). Large Language Models as General Purpose Intelligence Systems for Reasoning, Planning and Decision Making. American Journal of Artificial Intelligence and Neural Networks, 6(4), 45-72.

[10] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. IEEE Access, 12, 195974-195985.

[11] Zhang, F., & Yang, J. S. (2025). Learning Driven Decision Intelligence for Autonomous Driving Through Multimodal Understanding World Modeling and Policy Optimization. Frontiers in Artificial Intelligence Research, 2(3), 616-634.

[12] Wang, B., Wang, Z., Zhao, W., & Liu, Y. (2025). Network Fabric Simulation and Validation for Data Center Routing Convergence Under Large-Scale Failure Scenarios. Computer Science Bulletin, 8(01), 310-326.

[13] Nitsch, J., Itkina, M., Senanayake, R., Nieto, J., Schmidt, M., Siegwart, R., ... & Cadena, C. (2021, September). Out-of-distribution detection for automotive perception. In 2021 IEEE International Intelligent Transportation Systems Conference (ITSC) (pp. 2938-2943). IEEE.

[14] Fremont, D. J., Dreossi, T., Ghosh, S., Yue, X., Sangiovanni-Vincentelli, A. L., & Seshia, S. A. (2019, June). Scenic: a language for scenario specification and scene generation. In Proceedings of the 40th ACM SIGPLAN conference on programming language design and implementation (pp. 63-78).

[15] Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., & Geiger, A. (2022). Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE transactions on pattern analysis and machine intelligence, 45(11), 12878-12895.

[16] Chitta, K., Prakash, A., & Geiger, A. (2021). Neat: Neural attention fields for end-to-end autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 15793-15803).

[17] Feng, D., Harakeh, A., Waslander, S. L., & Dietmayer, K. (2021). A review and comparative study on probabilistic object detection in autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 23(8), 9961-9980.

[18] Malinin, A., & Gales, M. (2020). Uncertainty estimation in autoregressive structured prediction. arXiv preprint arXiv:2002.07650.

[19] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., ... & Dai, J. (2024). Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(3), 2020-2036.

[20] Liu, J., Wang, J., Chen, H., Guinness, J., Martin, R., & Kulkarni, C. S. (2019). Optimal Level Crossing Predictions for Electronic Prognostics. In AIAA Scitech 2019 Forum (p. 1962).

[21] Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. Applied Sciences, 14(20), 9213.

[22] Wei, Z., Sun, T., & Zhou, M. (2024). LIRL: Latent Imagination-Based Reinforcement Learning for Efficient Coverage Path Planning. Symmetry, 16(11), 1537.

[23] Zhang, S., Qiu, L., & Zeng, Z. (2026). Physics-Data Synergy in Structural Health Monitoring: A Multi-Scale Graph Contrastive Framework With Temperature-Adaptive Fusion. IEEE Access.

[24] Zeng, Z., Lin, H., Zhang, S., & Wang, B. (2026). Adaptive Robust Watermarking for Large Language Models via Dynamic Token Embedding Perturbation. IEEE Access, 14, 9319-9339.

[25] Qiu, L. (2025). Multi-Agent Reinforcement Learning for Coordinated Smart Grid and Building Energy Management Across Urban Communities. Computer Life, 13(3), 8-15.

[26] Zhao, W., Chen, T., Yang, J. S., & Qiu, L. (2026). AutoML-Pipeline: A RAG-enhanced code generation framework with pre-validation for cloud-native machine learning workflows. IEEE Access.

[27] Yang, Y., & Yang, J. (2026). Synthetic Data Meets Finance: Generative Models for Privacy Preserving Analytics. Journal of Banking and Financial Dynamics, 10(4), 1-8.

[28] Wang, Z., Shen, Z., Wang, B., & Shang, W. (2025). Modernizing Enterprise Analytics through Low-Code Automation and Cloud-Native Data Architectures. Asian Business Research Journal, 10(12), 20-33.

[29] Zhao, X., Sun, T., Ren, S., Yang, J., & Liu, Y. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment. Frontiers in Artificial Intelligence Research, 2(3), 501-520.

[30] Li, P., Liu, J., & Qiu, L. (2026). Deep Learning Methods for Demand Forecasting and Inventory Optimization in Modern Supply Chains. Asian Business Research Journal, 11(3), 21-29.

[31] Qiu, L. (2025). Reinforcement Learning Approaches for Intelligent Control of Smart Building Energy Systems with Real-Time Adaptation to Occupant Behavior and Weather Conditions. Journal of Computing and Electronic Information Management, 18(2), 32-37.

[32] Hanselmann, N., Renz, K., Chitta, K., Bhattacharyya, A., & Geiger, A. (2022, October). King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In European Conference on Computer Vision (pp. 335-352). Cham: Springer Nature Switzerland.

[33] Wu, P., Jia, X., Chen, L., Yan, J., Li, H., & Qiao, Y. (2022). Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. Advances in Neural Information Processing Systems, 35, 6119-6132.